Using Machine Learning to Identify Movie Genres through Online Movie Synopses

Jingcheng Wang
Department of Electronic and Electrical Engineering, University College London,
London WC1E 6BT, UK.

***Abstract*—Movie synopsis is important for audience to know about a movie within a short time. A good movie synopsis should reflect the genre, structure and main plot of a certain movie. The aim of this paper is to use machine learning to identify the genres of movie through movie synopses. The movies and corresponding synopses in database are downloaded from the Kaggle and ROTTEN TOMATOES websites. This study uses two supervised learning models (k-NN and SVM) and two deep learning models (CNN and RNN) to classify the genres of movie through movie synopses. Secondly, it tries to eliminate the interference by actively eliminating proper nouns. Finally, it compares and analyzes the performance of all models in different training sets. The result is that RNN with LSTM layer is the most suitable model for analyzing a large number of text for movie synopses, and the accuracy of judging movie genres is 80.5%. This study promotes the understanding of machine learning model selection in the adoption of movie genres classification based on movie synopses.**

Keywords—machine learning, k-Nearest Neighbors, Support Vector Machines, Convolutional Neural Network, Recurrent Neural Network, movie genres

I. INTRODUCTION

Watching online movie synopses has gradually become an important decision for consumers to judge whether the genres of movies meet their watching demands [1]. Compared with other official movie promotion or reviews, online movie synopses are often more objective and comprehensive [2, 3], because they provide the main plot of the movie. Therefore, the synopses play a vital role in movie ratings [2]. However, accessing useful information can be tedious and time-consuming. Watching synopses from different cultural backgrounds can also lead to misunderstanding [4]. At the same time, movie synopses focused on one aspect of the plot of the movie may cause consumers to misjudge the genre of the movie [5]. Thus, how to extract essential information from movie synopses is particularly important.

In recent years, there are also some methods to analyze movie synopses. Natural Language Processing (NLP) has developed rapidly with deep learning in recent years to explore the value of web pages [6, 7, 8, 9]. Therefore, it is a general trend to use programs to automatically process information on websites.

This paper investigates how well modern machine learning can do for movie synopses analysis. The aim of this paper is to classify the genres of movie through movie synopses. Some data cleaning and processing work are conducted, based on the 5000 movies downloaded from Kaggle and corresponding movie synopses from ROTTEN TOMATOES [10]. Then the data are divided into 90% training set and 10% test set, and then this research uses supervised learning models——k-Nearest Neighbors (k-NN) and Support Vector Machines (SVM) as well as convolutional neural network (CNN) and Recurrent Neural Network (RNN) in deep learning to extract the genre information of movies from movie synopses. Secondly, this paper eliminates proper nouns such as person names and place in synopses to verify whether the prediction results have been improved. Moreover, this study is going to compare the performance of these methods in different training sets.

II. RESEARCH DATA SETTINGS

A. Genre Dataset

This paper uses The Movie Database (TMDb) downloaded from Kaggle [10]. TMDb contains the data of 5000 movies, each of which has the attributes of movie name, movie genres, release time, duration, actor list, box office, etc. Since this research is only concerned with the classification of movies, it only focuses on the attributes of the movie name and genres of movies. Besides, there are hundreds of original genres classifications, which is a course of Dimensionality for subsequent processing [11, 12]. Therefore, this study uses 9 categories of genes to represent the whole genres as shown in Table 1.

Table 1. The categories of movie genres in this paper

Genre Categories	Genres	Number of samples
Action	Adventure, War and Military, Spy and Espionage Action, Martial Arts Action, Western Shoot ‘Em Up Action	925
Thriller	Conspiracy Thriller, Crime Thriller, Legal Thriller, Spy Thriller, Supernatural Thriller	442
Drama	Music Drama, Classic Western Drama	800

978-0-7381-1141-4/20/$31.00 ©2020 IEEE
DOI 10.1109/ITCA52113.2020.00008

Family	Kids, Animals	394
Romance	Historical Romance, Romantic Drama, Chick Flick, Paranormal Romance	796
Comedy	Slapstick Comedy, Screwball Comedy, Parody Comedy, Black Comedy	264
Horror	Zombie, Folk, Body, Found Footage	633
Animation	Hand drawn graphic animation, Virtual generated animation, Real people combined animation	295
Sci-Fi	Space Travel, Time Travel, Cerebral Science, Robot and Monster, Disaster and Alien Invasion	451

B. Synopsis Dataset

10000 synopses are downloaded from the ROTTEN TOMATOES website, every two of which correspond to the same movie in the movie genre dataset. Then, the two synopses of the same movie are simply added to form an entire information, which aims to make the feature attribute of individual sample more accurate [5].

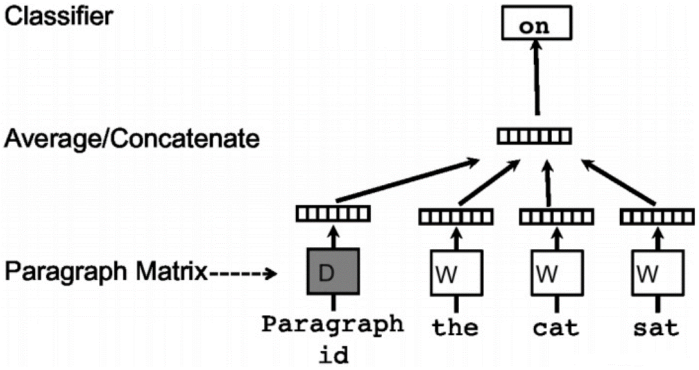

Figure 1. Distributed Memory Model of paragraph vectors

However, simple text cannot be read by computer, so text processing is needed [6]. When machine learning is used in text analysis, bag of words (BOW) is the most employed method to extract text [13]. As a result, Doc2Vec is adopted which is easy to use and understand, and is largely based on word2vec [2, 13, 14]. In doc2vec, one column of matrix D stands for one sentence. One column of the matrix W stands for one word. One word is chosen as the prediction word and the other is selected as the input word. The input layer consists of the sentence and word vector. A new vector X will be generated based one the input[9, 14]. This training method is called distributed memory model of paragraph vectors, as shown in Figure 1.

III. CLASSIFIERS CONFIGURATION

A. K-Nearest Neighbors

k-NN is one of the simplest methods in machine learning classification technology [15]. KNN based classifiers are decision tree, bagging and boost. Despite all this, KNN is still elected here because it is expected to observe how the simplest one will affect the results. KNN prediction is based on the idea that "similar things must be similar", which is realized by checking the labels of the nearest adjacent data points of test data points. Euclidean distance between two points x and x prime in d-dimensional Euclidean space is given.

$$\rho(x,x') = ||x - x'||_2^2 = \sqrt{\sum_{k=1}^{d} x_k - x'_k} \quad (1)$$

The task of this paper is to find in the range of K points, which kind of point is closest to test point, so the test point belongs to this category. In this formula, it is needed to pay attention to the selection of K, since if K is too small or to large, it will lead to over fitting, and high bias [15]. In this research, a set of K is tested and through cross validation to select the most appropriate K.

B. Support Vector Machine

SVM is the latest technology broadly used in statistical document classification [16, 17]. In principle, SVM is a binary sort model. Its basic strategy is to use a linear classifier to distinguish different types of data, and try to make the distance between the boundary and the data as far as possible [16]. Using karush Kuhn Tucker (KKT) condition and Lagrange formula, the decision boundary can be derived as follow.

$$d(X^T) = \sum_{i=1}^{l} y_i a_i X_i X^T + b_0 \quad (2)$$

The location of the test point can determine its classification. In this study, there are many kinds of species, so the research can make n-times binomial classification and then summarize them together. For the parameter adjustment step, in sklearn. kernel – Gaussian radial basis

function (RBF), the parameter gamma controls the number of support vector and C stands for the penalty coefficient. This study uses cross validation and grid search to find the suitable value.

C. CNN

With the development of deep learning theory, CNN is widely adopted in computer vision and text analysis [7, 18]. For normal CNN input, it needs to be processed by the embedding layer before the subsequent neural transmission to project the high-latitude, relatively sparse data to various dimensions and project it to a relatively low dimension [2]. Fortunately, the samples processed by doc2vec are embedded into the original embedding layer, which can significantly improve the accuracy and shorten the time.

In this study, the features are extracted and compressed by three convolution layers and max pooling layers. Max pooling is adopted instead of avg pooling because it only focuses on local features in semantics. Then, two full connected neural networks with dense layers are added to determine the classification of input. 'Relu' is used as an activator in the hidden layer while the activator of the output layer is 'softmax'. In addition, three dropout layers of 0.2 are added before the output layer to prevent overfitting.

D. RNN

RNN is a kind of recurrent neural network which takes sequence data as input. Its recursive process is carried out in the evolutionary direction of sequence, and all cyclic units are chained [19]. Because of its memory, RNN is very good at processing nonlinear feature sequences, so it is often used in NLP [20].

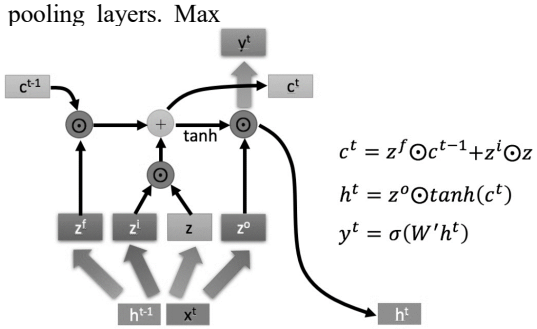

Figure 2. LSTM layer

Long short-term memory (LSTM) is used to improve the gradient disappearance and gradient explosion of RNN in iterative process [21]. In short, LSTM has a better robust in longer sequences than ordinary RNNs. The function of input gate is to determine the input of the current time step and update the system state to the internal state of the previous time step; the function of the forgetting gate is to update the internal state of the previous time step to the internal state of the current time step; the function of the output gate is to update the internal state of the current time step to the system state [19].

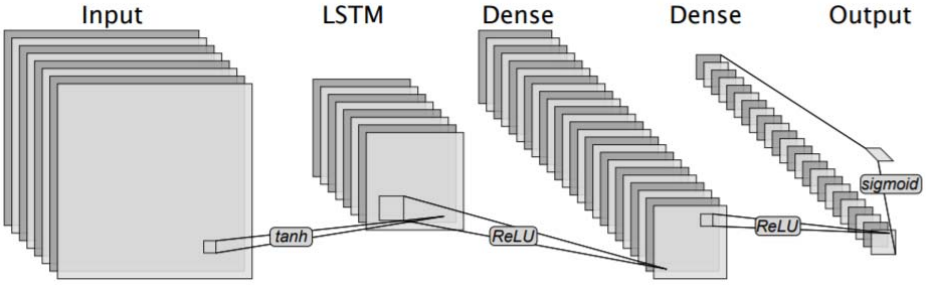

Figure 3. RNN architecture

In this study, it uses sigmoid cross entropy as the loss function. The other part is similar to CNN - The input is the embedding layer of Doc2Vec; 'Relu' is used as an activator in the hidden layer; Two dense layers are used to enable RNN to learn it hierarchical and compositional characteristics [19]. Since the output of the sigmoid colon is a probability value from 0 to 1, each neuron in this layer learns to recognize the probability of one of the observed features. Both RNN with LSTM layer and RNN without LSTM layer are trained to prove the importance of LSTM.

IV. RESULTS

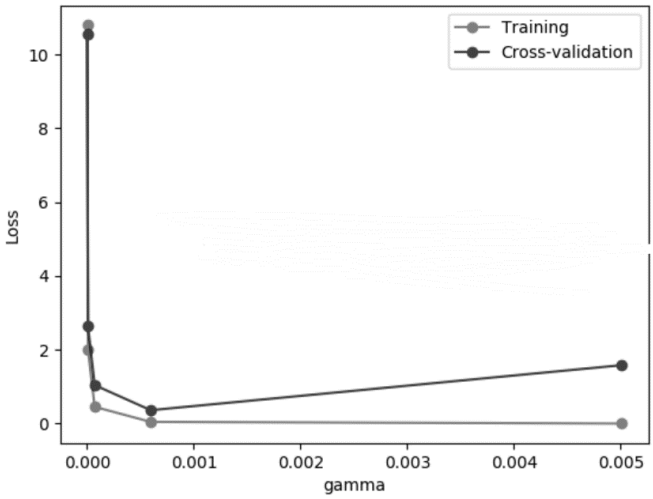

Figure 4 SVM parameter vs loss

Firstly, for KNN, through cross validation, the maximum accuracy can be got when k =[40, 60], k∈N. For SVM, when (C, gamma) = (10, 0.0006), the loss is minimum, as shown in figure 4.

Table 2. Test result with raw text

Machine Learning Model		Accuracy (epoch=20)	Accuracy (epoch=40)
Traditional Supervised Learning	k-NN	42.4	
	SVM	49.0	
Deep Learning	CNN	71.4	73.4
	RNN without LSTM layer	24.2	26.6
	RNN with LSTM layer	76.1	79.9

Table 3. Test result with 1/5 sample

Machine Learning Model		Accuracy with 1/5 sample (epoch=20)	Accuracy with 1/5 sample (epoch=40)
Traditional Supervised Learning	k-NN	42.3	
	SVM	49.0	
Deep Learning	CNN	39.5	47.1
	RNN without LSTM layer	23.9	24.8
	RNN with LSTM layer	44.4	52.2

Table 4. Test result for text without proper nouns

Machine Learning Model		Accuracy (epoch=20)	Accuracy (epoch=40)
Traditional Supervised Learning	k-NN	51.2	
	SVM	63.8	
Deep Learning	CNN	72.0	73.4
	RNN without LSTM layer	28.1	28.2
	RNN with LSTM layer	76.3	80.5

This paper trains the training set of the four machine learning methods (k-NN SVM, CNN, RNN with LSTM layer) with different epoch. Through modeling on the raw movie synopses, it is found that the highest accuracy is 79.9% (Table 3). So, this research takes the sampling method and reduce the training set to one fifth of the original, as a result, it can be found the highest accuracy reduces to 52.2%. After removing proper nouns, the highest accuracy can reach is 80.5% (as shown in Table 4).

V. DISCUSSION

A. Comparing the overall performance of different models

Compared with the traditional supervised learning and deep learning algorithms, it can be observed that deep learning always performs better. Especially for RNN, in different situations, the accuracy rate is more than 50%. It can be concluded that RNN is very suitable for movie genre classification through movie synopses. The effect of RNN alone is the worst, whose accuracy is only 24.1%.

For the deep learning model, RNN with LSTM layer has a poor effect because it is only suitable for very short sequences. Without proper forgetting method, the weight of the earliest data will become very low after iteration [19]. On the contrary, for movie synopses, it is clear that every sentence is equally important. Next, comparing CNN with RNN with LSTM layer, RNN with LSTM layer always leads CNN by about 10% because CNN cannot model the change of time series. However, the chronological order of sample appearance is very important for NLP. In addition, CNN is an extension of space, and neurons have convolutional features. Therefore, for large amounts of text data, the processing efficiency is low and there is no memory function. That is the reason why RNN with LSTM layer is better than CNN.

For traditional supervised learning, SVM has an overall high accuracy. k-NN prediction uses all training samples, which is a non-sparse model. While SVM only uses support vector, which is a sparse model. However, KNN is a kind of lazy learning without training process, which results in the need to reload all data for each run [15]. It takes more than 20 minutes to run by using 6G GPU. It can be seen that k-NN not only has low accuracy but also has low efficiency for large-scale data processing. Therefore, SVM outperforms KNN in various aspects.

B. Impact of data and epoch on results

When the number of sample in training set is reduced, the accuracy of deep learning method decreases obviously, while k-NN and SVM almost remain unchanged. In this case, SVM training results even surpass CNN. This is because for neural networks, it often needs a large number of sample data to learn in order to make the results better. However, for k-NN and SVM, it is only a simple classification problem and requires less sample size. With the increase of epoch, the accuracy is improved. However, this study does not continue to verify whether the loss will converge when epoch is equal, because this requires strong hardware and time to support.

Proper nouns are totally different in every movie, and for human beings themselves, they cannot understand the meaning of proper nouns for the first time. Therefore, proper nouns are deemed as interference items subjectively and all proper nouns are replaced by “it” or “them”. The results show that the improvement of k-NN and SVM is great, but the improvement of CNN and CNN is very small. This may be because in the continuous transmission, reverse transmission and drop of neural networks, the weights of these proper nouns do not interfere with the results, but for supervised learning, the elimination of these words will greatly reduce the classification interference. Besides, the purpose of adding deletions into neural network is to increase the noise, so as to train the anti-interference ability of the model. So generally speaking, for CNN and RNN, the original text can be processed, while KNN and SVM need further text processing.

VI. CONCLUSION

This study combines and processes a database of movie details from Kaggle and synopses from ROTTEN TOMATOES [10]. After that, four different machine learning models are used to infer the genres of movies according to the movie synopses and compare to values in the Kaggle database. In conclusion, if the number of samples in training set is limited, SVM or k-NN can be used but SVM is always better. Once the sample data of the training set increases, the training speed of SVM or k-NN increases rapidly, and the improvement of accuracy is less, so this is not applicable to the actual situation. Instead, CNN and RNN with LSTM layer in deep learning are more suitable to act as a text analysts for huge data. In this study, RNN with LSTM layer is the best model overall. Therefore, it can be sure that RNN with LSTM layer has a huge potential in semantic

analysis. In the future work, starring is also an important research direction. It is expected to improve the accuracy by analyzing the frequency of the stars participation in different synopses of movies and establishing the graph neural network between actors and actors.

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to my teachers for their guidance and suggestions. And I want to appreciate my family and friends for their care and love. Without their encouragement and help, I can not finish this paper.

REFERENCES

[1] Chen, Yubo, Liu, Yong & Zhang, Jurui, 2012. When do Third-Party Product Reviews Affect Firm Value and what can Firms Do? The Case of Media Critics and Professional Movie Reviews. Journal of marketing, 76(2), pp.116–134.

[2] Gabriel S Simoes, Jonatas Wehrmann, Rodrigo C Barros, and Duncan D Ruiz. 2016. Movie genre classification with convolutional neural networks. In Neural Networks (IJCNN), 2016 International Joint Conference on, pp. 259–266.

[3] Alexander L. Brown, Colin F. Camerer & Dan Lovallo, 2013. Estimating Structural Models of Equilibrium and Cognitive Hierarchy Thinking in the Field: The Case of Withheld Movie Critic Reviews. Management science, 59(3), pp.733–747.

[4] Koh, Noi Sian, Hu, Nan & Clemons, Eric K, 2010. Do online reviews reflect a product's true perceived quality? An investigation of online movie reviews across cultures. Electronic commerce research and applications, 9(5), pp.374–385.

[5] Manek, Asha S et al., 2016. Aspect term extraction for sentiment analysis in large movie reviews using Gini Index feature selection method and SVM classifier. World wide web (Bussum), 20(2), pp.135–154.

[6] Hardeniya, N. et al., 2016. Natural Language Processing: Python and NLTK / Hardeniya, Nitin. 1st ed.,

[7] LeCun, Yann, Bengio, Yoshua & Hinton, Geoffrey, 2015. Deep learning. Nature (London), 521(7553), pp.436–444.

[8] Yang, Zaoli et al., 2020. A decision-making algorithm for online shopping using deep-learning–based opinion pairs mining and q-rung orthopair fuzzy interaction Heronian mean operators. International journal of intelligent systems, 35(5), pp.783–825.

[9] Xue, Di et al., 2018. Deep learning-based personality recognition from text posts of online social networks. Applied intelligence (Dordrecht, Netherlands), 48(11), pp.4232–4246.

[10] The Movie Database, 2017. TMDB 5000 Movie Dataset. Retrieved from: https://www.kaggle.com/tmdb/tmdb-movie-metadata.

[11] Pestov, V., 2013. Is the k-NN classifier in high dimensions affected by the curse of dimensionality? Computers & mathematics with applications (1987), 65(10), pp.1427–1437.

[12] Chan, T.-H & Jiang, Shaofeng, 2018. Reducing Curse of Dimensionality. ACM Transactions on Algorithms (TALG), 14(1), pp.1–18.

[13] Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In Proceedings of the 31st International Conference on Machine Learning (ICML-14), pp. 1188–1196.

[14] Kim, Donghwa et al., 2019. Multi-co-training for document classification using various document representations: TF–IDF, LDA, and Doc2Vec. Information sciences, 477, pp.15–29.

[15] Subhrendu Gangopadhyay, Martyn Clark & Balaji Rajagopalan, 2005. Statistical downscaling using K-nearest neighbors. Water Resources Research, 41(2), pp. W02024-n/a.

[16] T. Joachims, "Training linear svms in linear time," in Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2006, pp. 217–226.

[17] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in European conference on machine learning. Springer, 1998, pp. 137–142.

[18] Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, L., Wang, G. and Cai, J., 2015. Recent advances in convolutional neural networks. arXiv preprint arXiv:1512.07108.

[19] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[20] 20. Jiang, Changhui et al., 2018. A MEMS IMU De-Noising Method Using Long Short Term Memory Recurrent Neural Networks (LSTM-RNN). Sensors (Basel, Switzerland), 18(10), p.3470.

[21] Goodfellow, I., Bengio, Y., Courville, A. Deep learning (Vol. 1): Cambridge: MIT Press, 2016: 367-415.